

Matrix Chebyshev expansion and its application to eigenspaces recovery

Nir Sharon^{* 1} and Yoel Shkolnisky^{†2}

¹The Program in Applied and Computational Mathematics, Princeton University, Princeton NJ, USA

²School of Mathematical sciences, Tel-Aviv University, Tel-Aviv, Israel

Abstract

We revisit the problem of lifting real functions to matrix functions via Chebyshev expansions. Given a real function f and a matrix with real spectrum A , we analyze the convergence of error bounds for the so-called matrix Chebyshev expansion of $f(A)$. The results shed light on the relation between the smoothness of f and the diagonalizability of the matrix A , and its effect on the approximation of $f(A)$ using the truncated matrix Chebyshev expansion. We present several numerical examples to illustrate our analysis and show an application of the matrix Chebyshev expansion for estimating eigenspaces of large matrices, associated with certain eigenvalues.

Key Words. Matrix functions, matrix Chebyshev expansion, convergence rates, inverse power method.

1 Introduction

Chebyshev polynomials are ubiquitous in applied math and engineering sciences¹. These polynomials arise as solutions of a Sturm-Liouville ODE and are used for numerous approximation methods, ranging from classical PDE methods [13], to modern methods for image denoising [15].

We focus on the problem of lifting a real function $f: \mathbb{R} \rightarrow \mathbb{R}$ to a matrix function $f: \mathcal{M}_k(\mathbb{R}) \rightarrow \mathcal{M}_k(\mathbb{R})$, where $\mathcal{M}_k(\mathbb{R})$ is the set of square real matrices of order k having real spectrum. When f is a polynomial, such a lifting is straightforward since addition and powers are well-defined for square matrices. When f is not a polynomial, there are several standard methods for defining the above lifting. If f is analytic with a convergence radius which is larger than the spectral radius of A , then the Taylor expansion $f(x) = \sum_{n=0}^{\infty} \alpha_n x^n$ yields $f(A) = \sum_{n=0}^{\infty} \alpha_n A^n$. If f is not analytic, it required that at least $f \in C^m$, where m is the size of the largest Jordan block of A , which allows to define $f(A)$ on each of the Jordan blocks of A . This latter approach has several equivalent definitions, see e.g. [8, Chapter 1]. In practice, most standard definitions suffer from numerical and computational drawbacks, which limits their applicability. This is the case, for example, with definitions that are based on the explicit form of the minimal polynomial of A or its Jordan form.

^{*}nsharon@math.princeton.edu

[†]yoelsh@post.tau.ac.il

¹“Chebyshev polynomials are everywhere dense in numerical analysis”. This quote by Philip Davis and George Forsythe is the opening sentence of [13].

Inspired by the numerical advantages of Chebyshev polynomials in representing and approximating scalar functions, we study the use of Chebyshev expansions for matrix functions. The idea of evaluating a matrix function by its Chebyshev expansion is not new. In [21] it is used in spectral methods for solving PDEs. In this context, the Chebyshev polynomials are also examined as a special case of ultraspherical polynomials [4], and Feber polynomials [20]. Expansion in the Feber polynomials (for complex spectra) is also studied for matrices in [16]. Chebyshev polynomials are also widely used for pseudospectral methods, see [22] and reference therein. In all the above papers it is assumed that f is a smooth function. One important analytic function is the exponential, used naturally for solving differential equations. In [1] Chebyshev polynomials are proved to be an effective alternative to Krylov techniques for calculating $\exp(A)v$, for a given vector v . Another application of matrix Chebyshev polynomials is in representing the best matrix 2-norm approximation for analytic functions, over the space of polynomials of a fixed degree [12]. We can also find matrix Chebyshev polynomials in calculating matrix functions of symmetric matrices [6], computing square roots of the covariance matrix of Gaussian random fields [3], facilitating the estimation of autoregressive models [17], and in the construction of image denoising operators [15].

Our contribution

In this paper we generalize the study of matrix Chebyshev expansions to the case where the matrix is not necessarily diagonalizable, and where the function is not necessarily analytic. By making these assumptions we reveal a trade-off between two factors; “how much” the matrix is far from being diagonalizable, as expressed by the size of the largest Jordan block in the Jordan form of the matrix, and “how smooth is the function”. Specifically, as the size of the largest Jordan blocks increases (“less diagonalizable”) the smoothness required from the function in order to guarantee the convergence of the matrix Chebyshev expansion is higher.

The contribution of the current paper consists of both a theoretical analysis and a practical application. In the analysis part we examine convergence issues and prove convergence rates. The convergence rate of the matrix Chebyshev expansion is important since it is crucial in determining how many coefficients one has to use to approximate $f(A)$ to a prescribed accuracy using a truncated Chebyshev expansion. Thus, the convergence rate directly affects the efficiency of algorithms that are based on the matrix Chebyshev expansion. Since Chebyshev polynomial are naturally defined on $[-1, 1]$, we divide our analysis to two cases: a case where there is no a priori information regarding the distribution of the eigenvalues over $[-1, 1]$, and a case where all non-semisimple eigenvalues are concentrated inside $[-1, 1]$. As expected, the latter case implies less restrictions on the smoothness of the function. In particular, if we denote by m the size of the largest Jordan block of A , then the convergence of the matrix Chebyshev expansion is guaranteed in the first case for any $f \in C^{2m-2}$ where $f^{(2m-1)}$ is of bounded variation, while in the second case it is true for any $f \in C^{m-1}$ where $f^{(m)}$ is of bounded variation. As for convergence rates, we summarize our results for the different cases in Table 1. Each row of the table presents the requirements, from both the matrix and the function, for an expansion of length N to have an error bound of $cN^{-\ell}$ ($\ell > 1$). The constant c is independent of N and ℓ but does depend on other factors as mentioned under remarks. We note two properties from the table. First, for matrices of concentrated spectrum, the smoothness assumptions on f are weaker. Second, one can guarantee a constant which is independent of the size of the matrix k by imposing some extra regularity on the function f . Further conclusions and

Spectrum of A	The function	Remarks
A is a δ -condense matrix	$f \in C^{m+\ell-2}$ and $f^{(m+\ell-1)}$ is of bounded variation	All non-semisimple eigenvalues of A are in $[-1+\delta, 1-\delta]$ c depends on k, f, m
	$f \in C^{2m+\ell-3}$ and $f^{(2m+\ell-2)}$ is of bounded variation	c depends on k, f, m
A is a δ -condense matrix	$f \in C^{m+2\ell-1}$ and $f^{(m+2\ell)}$ is of bounded variation	All non-semisimple eigenvalues of A are in $[-1+\delta, 1-\delta]$ c depends on f, m
	$f \in C^{2m+2\ell}$ and $f^{(2m+2\ell+1)}$ is of bounded variation	c depends on f and m

Table 1: Table of conditions to establish a convergence rate of order $N^{-\ell}$, with a constant c , for a truncated matrix Chebyshev expansion of length N . The matrix is of size $k \times k$ and with m as the size of its largest Jordan Block

their proofs appear in the text.

In the application part of the paper we demonstrate the efficiency of matrix Chebyshev expansions for extracting eigenvectors associated with a specific predefined segment of the spectrum. This problem is usually addressed by inverse power iterations, where one has to solve a linear system at each iteration, see e.g., [7, Chapter 8]. Instead, we show that by designing an appropriate filter function, we can apply a truncated Chebyshev expansion and retain fast and accurate results in just a few iterations. This method provides a significant benefit as the size of the matrix increases.

The structure of the paper

The paper is organized as follows. In Section 2 we present our notation and the required mathematical background. Section 3 introduces the matrix Chebyshev expansion and our main theoretical results. These results include the conditions for convergence of a matrix Chebyshev expansion to the matrix function and bounds on the convergence rates. We illustrate numerically some of our theoretical results in Section 4 and demonstrate the application of the matrix Chebyshev expansion for the problem of subspace recovery in Section 5.

2 Background and notation

Chebyshev polynomials of the first kind of degree n are defined as

$$T_n(x) = \cos(n \arccos(x)), \quad x \in [-1, 1], \quad n = 0, 1, 2, \dots \quad (1)$$

These polynomials are solutions of the Sturm-Liouville ordinary differential equation

$$(1 - x^2)y'' - xy' + n^2y = 0 \quad (2)$$

and consequently satisfy the three term recursion

$$T_n(x) = 2xT_{n-1} - T_{n-2}, \quad n = 2, 3, \dots \quad (3)$$

with $T_0(x) = 1$ and $T_1(x) = x$. Therefore, Chebyshev polynomials form an orthogonal basis for $L_2([-1, 1])$ with respect to the inner product

$$\langle f, g \rangle_T = \frac{2}{\pi} \int_{-1}^1 \frac{f(t)g(t)}{\sqrt{1-t^2}} dt. \quad (4)$$

The Chebyshev expansion for any f with finite norm with respect to (4) is

$$f(x) \sim \sum_{n=0}^{\infty} \alpha_n[f] T_n(x), \quad \alpha_n[f] = \langle f, T_n \rangle_T. \quad (5)$$

where the dashed sum $\left(\sum'\right)$ denotes that the first term is halved. The truncated Chebyshev expansion is defined as

$$\mathcal{S}_N(f)(x) = \sum_{n=0}^N \alpha_n[f] T_n(x). \quad (6)$$

$\mathcal{S}_N(f)(x)$ provides a polynomial approximation which is the best least squares approximation with respect to the induced norm $\|f\|_T = \sqrt{\langle f, f \rangle_T}$. Remarkably, this least squares approximation is close to the best minimax polynomial approximation, measured by the maximum norm $\|f\|_\infty = \max_{x \in [-1, 1]} |f(x)|$. The following theorem was established by Bernstein [2] back in 1918 (and was known even before, see references therein).

Theorem 2.1 (Bernstein). *For any $f \in C([-1, 1])$ and $N \geq 0$*

$$\|f(x) - \mathcal{S}_N(f)(x)\|_\infty \leq \lambda_N \|f(x) - p_N^*(x)\|_\infty,$$

where p_N^* is the unique best minimax polynomial approximation of degree N , and λ_N behaves asymptotically as $\log(N)$ for large N .

The constant λ_N is the Lebesgue constant, and a formula for its exact value has been derived in [18] (and in particular, it is less than 6 for $N < 1000$). Note that on the boundaries of $[-1, 1]$ we only consider one-sided continuity (and derivatives when required).

Define the total variation (TV) norm by

$$\|g\|_{TV} = \|g'\|_1 = \int_{-1}^1 |g'(t)| dt.$$

Note that this norm can be written in a distributional form instead of the derivative form and thus we say that a function has a bounded variation if its TV norm is finite (whether continuous or not, e.g., the step function). Using the TV norm it was recently shown that for a smooth f , the uniform error $\|f(x) - \mathcal{S}_N(f)(x)\|_\infty$ decays rapidly [23, Chapter 7]

Theorem 2.2 (Trefethen). *Let $f \in C^{m-1}([-1, 1])$, where $f^{(m-1)}$ is absolutely continuous such that*

$\|f^{(m)}\|_{TV} < \infty$. Then, for each $N > m$, the Chebyshev coefficients satisfy

$$\alpha_N[f] \leq \frac{C \cdot m}{N(N-1) \cdots (N-m)} \leq (C \cdot m) \frac{1}{(N-m)^{m+1}}$$

and moreover

$$\|f(x) - \mathcal{S}_N(f)(x)\|_\infty \leq C \frac{1}{(N-m)^m},$$

where $C = C(f, m) = \frac{2}{\pi m} \|f^{(m)}\|_{TV}$.

The trigonometric form (1) implies that $T_n(\cos(\theta)) = \cos(n\theta)$, $n = 0, 1, 2, \dots$, and reveals that the Chebyshev expansion of a given $f(x)$, $x \in [-1, 1]$ coincides with the Fourier cosine series $g(\theta) = \sum_{n=0}^{\infty} \alpha_n[f] \cos(n\theta)$ where $g(\theta) = f(\cos(\theta))$ with $\theta \in [0, \pi]$. The argument $\cos(\theta)$ in f makes g even and periodic in θ , which implies that $g^{(j)}(\pm\pi) = 0$. Thus, no periodicity is required from f nor from its derivatives to ensure that g is differentiable and its Fourier series can be differentiated term by term. We conclude the above discussion with the next lemma.

Lemma 2.3. *Let $f \in C^{2m-2}([-1, 1])$ be such that $f^{(2m-2)}$ is absolutely continuous with $f^{(2m-1)}$ of bounded variation, $m \in \mathbb{N}$. Then,*

1. *The series*

$$\sum_{n=0}^{\infty} \alpha_n[f] T_n^{(j)}(x), \quad j = 0, 1, \dots, m-1, \quad x \in [-1, 1],$$

is absolutely convergent.

2. *One can differentiate the Chebyshev expansion term-by-term to have*

$$f^{(j)}(x) = \sum_{n=0}^{\infty} \alpha_n[f] T_n^{(j)}(x), \quad j = 0, 1, \dots, m, \quad x \in [-1, 1].$$

The proof of Lemma 2.3 is given for completeness in Appendix A.1.

3 Matrix Chebyshev expansion

This section presents the theoretical results of the paper; we define a matrix function via its Chebyshev expansion, discuss its convergence, and derive its convergence rate. To this end, unless otherwise stated, we assume a given matrix norm $\|\cdot\|$ that satisfies the sub-multiplicative property $\|XY\| \leq \|X\| \|Y\|$. Equipped with $\|\cdot\|$, we restrict the convergence analysis to absolute convergence (in norm). Namely, given a series of matrices $\{X_n\}_{n \in \mathbb{N}} \subset \mathcal{M}_k(\mathbb{R})$, we consider the convergence of $\|X_n\|$ as n tends to infinity. Note that since all matrix norms are equivalent, absolute convergence in one norm implies the absolute convergence in any other norm. For more details about convergence of matrix series and matrix norms, we refer the reader to [10, Chapter 5]. A short introduction to matrix functions and their basic definitions is provided for completeness in Appendix B.

3.1 Definition and convergence

In the sequel, we denote by $\lambda_1, \dots, \lambda_k \in \mathbb{R}$ the eigenvalues of a given matrix $A \in \mathcal{M}_k(\mathbb{R})$ and by $\rho(A) = \max_{i=1, \dots, k} |\lambda_i|$ the spectral radius of A . Since we discuss Chebyshev polynomials, we assume $\rho(A) \leq 1$ and that f is a scalar function defined on $[-1, 1]$. Also, we denote by $\|X\|_F = \sqrt{\text{tr}(XX^T)}$ the Frobenius norm of X .

We begin by defining the matrix Chebyshev expansion.

Definition 1 (Matrix Chebyshev expansion). *Given a function f on $[-1, 1]$, we define the partial sum*

$$\mathcal{S}_N(f)(A) = \sum_{n=0}^N \alpha_n[f] T_n(A),$$

where $\alpha_n[f]$ is given in (5). In addition, if the limit $\mathcal{S}_\infty(f)(A)$ is absolutely convergent with respect to any matrix norm $\|\cdot\|$, we define $\mathcal{S}_\infty(f)(A)$ as the matrix Chebyshev expansion,

$$\mathcal{S}_\infty(f)(A) = \lim_{N \rightarrow \infty} \mathcal{S}_N(f)(A). \quad (7)$$

In this subsection, we address two fundamental questions: how can one guarantee the (absolute) convergence of $\mathcal{S}_\infty(f)(A)$? And, does $\mathcal{S}_\infty(f)(A)$ coincide with $f(A)$ when using the standard definitions of matrix functions (see Appendix B)?

A basic property of polynomials, and specifically the Chebyshev polynomial T_n , is that they preserve matrix similarity. Namely, $A = P^{-1}BP$ implies $T_n(A) = P^{-1}T_n(B)P$. Therefore, for a diagonalizable matrix A , evaluating $T_n(A)$ reduces to apply $T_n(x)$ on the eigenvalues of A . Thus, for such matrices, the convergence of the Chebyshev expansion of a scalar function is inherited by its matrix version. This is summarized in the following theorem.

Corollary 3.1. *Let f be a function on $[-1, 1]$ having an absolutely convergent Chebyshev series. Then, for any diagonalizable matrix A , the matrix Chebyshev expansion of $f(A)$ is convergent.*

Proof. The diagonalizability means that $A = Q^{-1}DQ$, where $D = \text{diag}(\lambda_1, \dots, \lambda_k)$. Therefore, $\|T_n(A)\| \leq \|Q^{-1}\| \|T_n(D)\| \|Q\|$ and also $|T_n(\lambda_i)| \leq 1$, $1 \leq i \leq k$, see e.g., (1), which implies that $\|T_n(D)\|_F \leq \sqrt{k}$. Thus, we get

$$\sum_{n=0}^{\infty} \|\alpha_n[f] T_n(A)\|_F \leq \|Q^{-1}\|_F \|Q\|_F \sqrt{k} \sum_{n=0}^{\infty} |\alpha_n[f]| < \infty.$$

□

For a general matrix $A \in \mathcal{M}_k(\mathbb{R})$, we can try a similar argument but with the Jordan form $A = Z^{-1}JZ$, where $J = \text{diag}(J_1, \dots, J_p)$ is a diagonal block matrix with the blocks J_1, \dots, J_p on its diagonal. Since applying a polynomial on a block diagonal matrix reduces to applying it to each block separately we get that,

$$T_n(A) = Z^{-1} \text{diag}(T_n(J_1), \dots, T_n(J_p)) Z, \quad n \in \mathbb{N}. \quad (8)$$

Thus, in order to understand $T_n(A)$, it is s to understand $T_n(J_i)$, where J_i is a Jordan block of order which we denote by k_i (see also the explicit form in (33)).

Lemma 3.2. Let $p(x) = \sum_{n=0}^{\ell} a_n x^n$ be a polynomial of degree ℓ . Then

$$p(J_i) = \begin{pmatrix} p(\lambda_i) & p'(\lambda_i) & \cdots & \frac{p^{(k_i-1)}(\lambda_i)}{(k_i-1)!} \\ & p(\lambda_i) & \ddots & \vdots \\ & & \ddots & p'(\lambda_i) \\ & & & p(\lambda_i) \end{pmatrix}.$$

where J_i is the $k_i \times k_i$ Jordan block, as in (33).

The proof of Lemma 3.2 is given in Appendix A.2. Obviously, this result must agree with standard definitions of matrix functions, see e.g., (34) of Definition 5 in Appendix B.

The explicit form of $T_n(J_i)$, derived from Lemma 3.2, gives rise to the next convergence result. The only additional fact we need for generalizing Corollary 3.1 to non-diagonalizable matrices is a bound on the derivatives of the Chebyshev polynomials. In [19, Chapter 1.5] it is proven that for any $j \in \mathbb{N}$

$$\left| T_n^{(j)}(x) \right| \leq \left| T_n^{(j)}(1) \right| = \frac{n^2(n^2-1) \cdots (n^2-(j-1)^2)}{(2j-1) \cdots 5 \cdot 3 \cdot 1} \leq \frac{n^{2j}}{(2j-1)!}. \quad (9)$$

Therefore we have,

Theorem 3.3. Let $A \in \mathcal{M}_k(\mathbb{R})$ be a matrix and let $m \geq 1$ be the size of the largest Jordan block in the Jordan form of A . Assume that $f \in C^{2m-2}([-1, 1])$ such that $f^{(2m-2)}$ is absolutely continuous and $f^{(2m-1)}$ is of bounded variation. Then, the Chebyshev expansion $\mathcal{S}_{\infty}(f)(A)$ is absolutely convergent.

Proof. It is convenient to prove the theorem using the ℓ_1 -induced matrix norm, $\|X\|_1 = \max_i \sum_j |X_{i,j}|$ whereby we get

$$\|T_n(A)\|_1 \leq \|Z^{-1}\|_1 \left(\max_{i=1, \dots, p} \|T_n(J_i)\|_1 \right) \|Z\|_1, \quad n \in \mathbb{N}. \quad (10)$$

By the explicit form given in Lemma 3.2, for any Jordan block J_i of order k_i

$$\|T_n(J_i)\|_1 = \sum_{j=1}^{k_i} \left| \frac{1}{(j-1)!} \frac{d^{j-1}}{dx^{j-1}} T_n(\lambda_i) \right|, \quad n \in \mathbb{N}. \quad (11)$$

Therefore (11) is bounded by

$$\|T_n(J_i)\|_1 \leq 1 + \sum_{j=1}^{k_i-1} \frac{1}{j!} \left| T_n^{(j)}(\lambda_i) \right| \leq n^{2k_i-2} \left(1 + \sum_{j=0}^{\infty} 2^{-j} \right) \leq 3n^{2k_i-2}, \quad n \in \mathbb{N}. \quad (12)$$

Recall that $m = \max_{i=1, \dots, p} k_i$ and combine (10) and (12) to get

$$\|T_n(A)\|_1 \leq \|Z^{-1}\|_1 \|Z\|_1 \max_{i=1, \dots, p} 3n^{2k_i-2} \leq C_A n^{2m-2}, \quad (13)$$

where $C_A = 3\|Z^{-1}\|_1\|Z\|_1$ is a constant that depends solely on the Jordan form of A and is independent of n . By Theorem 2.2, there exists C_f such that $|\alpha_n[f]| \leq C_f n^{-2m}$ for large enough n . Therefore, we

have that

$$\sum_{n=n^*}^{\infty} \|\alpha_n[f]T_n(A)\|_1 = \sum_{n=n^*}^{\infty} |\alpha_n[f]| \|T_n(A)\|_1 \leq C_f C_A \sum_{n=n^*}^{\infty} n^{-2} < \infty, \quad (14)$$

and so $\mathcal{S}_{\infty}(f)(A)$ of (7) is absolutely convergent. \square

We next show that $\mathcal{S}_{\infty}(f)(A)$ coincides with the standard definitions for $f(A)$. To answer the second question given after Definition 1 we have,

Theorem 3.4. *Let $A \in \mathcal{M}_k(\mathbb{R})$ and let f satisfy the conditions for an absolutely convergent matrix Chebyshev expansion. Then, $\mathcal{S}_{\infty}(f)(A) = f(A)$, where $f(A)$ is one of the standard definitions for lifting a scalar function to matrices, as appear in Appendix B.*

Proof. We prove that matrix Chebyshev expansion is equivalent to Definition 5 in Appendix B since all standard definitions for lifting a scalar function to matrices are equivalent. We use the notation of the proof of Theorem 3.3 and by (8) deduce that it is enough to show that the definitions coincide on the Jordan blocks of A . Note that in the special case where the matrix is normal, as in Corollary 3.1, the definitions are equivalent. This is true due to the absolute convergence which implies pointwise convergence of the Chebyshev series for numbers, together with the fact that the function f is evaluated element-wise on the diagonal form.

Denote by m the size of the largest Jordan block of A . An element in the matrix $\mathcal{S}_{\infty}(f)(J_i)$, corresponding to the Jordan block J_i and the eigenvalue λ_i , is given for every $1 \leq p \leq j \leq m$ by

$$(\mathcal{S}_{\infty}(f)(J_i))_{p,j} = \left(\sum_{n=0}^{\infty} \alpha_n[f] T_n(J_i) \right)_{p,j} = \sum_{n=0}^{\infty} \alpha_n[f] (T_n(J_i))_{p,j}. \quad (15)$$

Note that the absolute convergence of the matrix Chebyshev expansion implies the absolute and uniform convergence of the scalar Chebyshev expansion of f . Thus, one can differentiate the Chebyshev series of f term-by-term to have $f^{(j)}$, $j = 0, \dots, m-1$ (see Lemma 2.3). Since Lemma 3.2 states that

$$(T_n(J))_{p,j} = \frac{1}{(j-p)!} \frac{d^{j-p}}{dx^{j-p}} T_n(\lambda_i),$$

we get that for $1 \leq p \leq j \leq k_i \leq m$

$$(\mathcal{S}_{\infty}(f)(J_i))_{p,j} = \frac{1}{(j-p)!} \sum_{n=0}^{\infty} \alpha_n[f] \frac{d^{j-p}}{dx^{j-p}} T_n(\lambda_i) = \frac{1}{(j-p)!} f^{(j-p)}(\lambda_i),$$

which coincides with (34) of Definition 5, as required. \square

3.2 Relaxing the smoothness requirements

The proof of Theorem 3.3 relies upon the bound (9) that can be improved for cases where the eigenvalues of A lie strictly inside the interval $[-1, 1]$. We start by examining the first derivative of $T_n(x)$, which by

(1) is $T'_n(x) = n \frac{\sin(n\theta)}{\sin(\theta)}$ for $x = \cos(\theta)$. Therefore, by [11]

$$|T'_n(x)| \leq \frac{n}{\sqrt{1-x^2}}, \quad |x| < 1, \quad n = 0, 1, 2, \dots \quad (16)$$

The factor $\frac{1}{\sqrt{1-x^2}}$ increases near the end points of $[-1, 1]$ and ultimately forces us to use the global bound of n^2 , see e.g., [19, Chapter 1]. However, for a segment $I_\delta = [-1 + \delta, 1 - \delta]$, with a fixed $0 < \delta < 1$ we can use a linear bound νn with

$$\nu = \frac{1}{\sqrt{2\delta(1-\delta)}}. \quad (17)$$

For higher derivatives, one can use (2) to derive [19, Chapter 1, p. 32–33]

$$(1-x^2)T_n^{(k+1)}(x) - (2k-1)xT_n^{(k)}(x) + (n^2 - (k-1)^2)T_n^{(k-1)}(x) = 0, \quad 1 \leq k \leq n.$$

Thus, we get

$$|T_n^{(k+1)}(x)| \leq \frac{(2k-1)|x|}{(1-x^2)}|T_n^{(k)}(x)| + \frac{n^2}{1-x^2}|T_n^{(k-1)}(x)|, \quad |x| < 1, \quad 1 \leq k \leq n. \quad (18)$$

The latter leads to the following bound on higher order derivatives of $T_n(x)$.

Lemma 3.5. *Let $0 < \delta < 1$ and $1 \leq k \leq n$. Then, the k^{th} derivative of $T_n(x)$ satisfies*

$$|T_n^{(k)}(x)| \leq c_k n^k, \quad |x| \leq \delta,$$

where $c_k = c_k(\nu, k)$ is a constant independent of n , and ν is given by (17).

Proof. We prove the lemma by induction on the order k of the derivative. For the basis of our induction we have $|T_n^{(0)}(x)| \leq 1 = (\nu n)^0$. For $k = 1$ the lemma holds due to (16) and the fact that $x \in I_\delta$. Now, assume the claim is true for any j that satisfies $0 \leq j \leq k < n$, for a fixed k . Then, for $k+1$ we get by (18) a leading order term n^{k+1} from the leading order of $|T_n^{(k-1)}(x)|$. Namely, $|T_n^{(k+1)}(x)| \leq n^2 \nu^2 (\nu n)^{k-1} + c n^k$. It is understood from (18) that the constant c of the lower order term (LOT) depend on k and ν but not on n . \square

Remark 1. *We did not elaborate on the LOTs of the bound on $T_n^{(k+1)}(x)$, which appear in the proof as $c n^k$. For example, with further induction, one can derive that the constant in front of the n^k term is bounded by $\frac{(k+1)k}{2} \nu^{k+2}$. The proof is given for completeness in Appendix A.3. Nevertheless, the explicit form of the constants of the LOTs are less important for us as we aim to pose convergence results, which are typically stated for large n . Thus, the explicit forms of the LOTs are omitted.*

Lemma 3.5 shows that in many cases the bound of n^2 for $T'_n(x)$ is very pessimistic and so are the conditions of Theorem 3.3 above. We therefore define the class of matrices for which we can relax the smoothness requirements from the function.

Definition 2 (δ -condense matrix). *Let $0 < \delta < 1$. We call a matrix $A \in \mathcal{M}_k(\mathbb{R})$ with $\rho(A) \leq 1$ a δ -condense matrix if any eigenvalue in $[-1, -\delta] \cap [\delta, 1]$ is semisimple (its algebraic multiplicity is equal to its geometric multiplicity).*

Having the above definition, we get the next convergence result.

Corollary 3.6. *Let $A \in \mathcal{M}_k(\mathbb{R})$ be a δ -condense matrix. Denote by $m \geq 1$ the size of the largest Jordan block in the Jordan form of A , and assume that $f \in C^{m-1}([-1, 1])$ such that $f^{(m-1)}$ is absolutely continuous with $f^{(m)}$ of bounded variation. Then, the Chebyshev expansion $\mathcal{S}_\infty(f)(A)$ is absolutely convergent.*

Proof. We follow the proof of Theorem 3.3 and modify it as described next. First, using Lemma 3.5 we replace (12) with

$$\|T_n(J_i)\|_1 \leq 1 + \sum_{j=1}^{k_i-1} \frac{1}{(j)!} |T_n^{(j)}(\lambda_i)| \leq 1 + \sum_{j=1}^{k_i-1} \frac{1}{(j)!} c_j n^j \leq \widetilde{c}_{k_i} n^{k_i-1},$$

for a constant \widetilde{c}_{k_i} that depends on k_i and ν but is independent of n . Since by definition $m = \max_{i=1, \dots, p} k_i$, we can bound

$$\|T_n(A)\|_1 \leq K n^{m-1}, \quad K = \max_{i=1, \dots, p} \widetilde{c}_{k_i} \|Z^{-1}\|_1 \|Z^{-1}\|_1.$$

Here K is a constant that depends only on ν and on the Jordan form of A . By Theorem 2.2, the assumptions on f ensure that (14) still holds. \square

Remark 2. *Using the same arguments above, we can also update the conditions of Lemma 2.3 to $f \in C^{m-1}([-1, 1])$ such that $f^{(m-1)}$ is absolutely continuous and $f^{(m)}$ is of bounded variation, for cases where $x \in [-\delta, \delta]$ with a positive $\delta < 1$.*

3.3 The truncated matrix Chebyshev expansion

The convergence rates of the Chebyshev expansion for scalar functions is well-studied (see Section 2) and we wish to extend some of the results to the matrix case. In particular, we generalize Theorem 2.2 to the case of matrix functions, in light of the additional factors we mentioned in the previous section.

In the case of a diagonalizable matrix $A = Q^{-1}DQ$ we bound the convergence rate of $\mathcal{S}_N(f)(A)$ based on the convergence rate of the scalar function f as

$$\|\mathcal{S}_N(f)(A) - f(A)\| \leq \|Q^{-1}\| \|\mathcal{S}_N(f)(D) - f(D)\| \|Q\|,$$

where the Chebyshev expansion on the diagonal matrix D acts element-wise. Next, we focus on more general cases where the matrices are allowed to be non-diagonalizable. In the following result we show how the smoothness of f is translated to convergence rate. Specifically, given extra ℓ derivatives over the required for convergence in Theorem 3.3 implies extra ℓ powers of error decay.

Theorem 3.7. *Let $A \in \mathcal{M}_k(\mathbb{R})$ and let $m \geq 1$ be the size of the largest Jordan block in the Jordan form of A . Assume $f \in C^{2m-2+\ell}([-1, 1])$ such that $f^{(2m-2+\ell)}$ is absolutely continuous and $f^{(2m-1+\ell)}$ is of bounded variation. Then, for a given matrix norm $\|\cdot\|$ we have*

$$\|\mathcal{S}_N(f)(A) - f(A)\| \leq C \frac{1}{(N - (2m - 1 + \ell))^{\ell+1}}, \quad N > 2m - 1 + \ell,$$

where the constant $C = C(f, \ell, m, \|\cdot\|)$ is independent of N .

Proof. Since by Theorem 3.3 and Theorem 3.4 $\mathcal{S}_\infty(f)(A)$ converges in norm to $f(A)$ we have

$$\|\mathcal{S}_N(f)(A) - f(A)\| = \left\| \sum_{n=N+1}^{\infty} \alpha_n[f] T_n(A) \right\|,$$

which can be bounded by $\sum_{n=N+1}^{\infty} |\alpha_n[f]| \|T_n(A)\|$. By the equivalence of norms we have that $\|\cdot\| \leq C_{\text{norm}} \|\cdot\|_1$ for some constant C_{norm} , and so by (13)

$$\|T_n(A)\| \leq C_{\text{norm}} C_A n^{2m-2}. \quad (19)$$

The assumptions on f and Theorem 2.2 imply that

$$|\alpha_n[f]| \leq \frac{2V}{\pi} \frac{1}{n(n-1) \cdots (n-(2m-1+\ell))}, \quad n > 2m-1+\ell, \quad (20)$$

with $V = \|f^{(2m-1+\ell)}\|_{TV}$. Note that $\frac{n^{2m-2}}{n(n-1) \cdots (n-(2m-1+\ell))}$ can be simplified as

$$\frac{n^{2m-3}}{(n-1) \cdots (n-(2m-3))} \frac{1}{(n-(2m-2)) \cdots (n-(2m-1+\ell))},$$

which is bounded by

$$\left(\frac{n}{n-(2m-3)} \right)^{2m-3} \left(\frac{1}{n-(2m-1+\ell)} \right)^{\ell+2}.$$

For $n > 2m-1+\ell$ we have

$$\frac{n}{n-(2m-3)} = 1 + \frac{2m-3}{n-(2m-3)} \leq 1 + \frac{2m-3}{\ell+2}.$$

Therefore, we denote $C = C_{\text{norm}} C_A \left(\frac{2V}{\pi} \right) \left(1 + \frac{2m-3}{\ell+2} \right)^{2m-3}$ and combine (19) and (20) to have

$$\begin{aligned} \sum_{n=N+1}^{\infty} |\alpha_n[f]| \|T_n(A)\| &\leq C \sum_{n=N+1}^{\infty} \left(\frac{1}{n-(2m-1+\ell)} \right)^{\ell+2} \\ &\leq C \int_N^{\infty} \frac{dx}{(x-(2m-1+\ell))^{\ell+2}} \\ &= \left(\frac{C}{\ell+2} \right) \frac{1}{(N-(2m-1+\ell))^{\ell+1}}. \end{aligned}$$

□

The last proof can be modified to the case of δ -condense matrices. In particular, (19) turns to $\|T_n(A)\| \leq C_{\text{norm}} C_A c_{m-1} n^{m-1}$. Therefore, the requirements from f can be weakened and the remaining calculations of the proof hold with $m-1$ replacing $2m-2$. We summarize it in the following.

Corollary 3.8. *Let $A \in \mathcal{M}_k(\mathbb{R})$ be a δ -condense matrix and let $m > 1$ be the size of the largest Jordan block in its Jordan form. Assume $f \in C^{m-1+\ell}([-1, 1])$ such that $f^{(m-1+\ell)}$ is absolutely continuous and*

$f^{(m+\ell)}$ is of bounded variation. Then, for a given matrix norm we have

$$\|\mathcal{S}_N(f)(A) - f(A)\| \leq C \frac{1}{(N - (m + \ell))^{\ell+1}}, \quad N > m + \ell,$$

where the constant $C = C(f, \ell, m, \|\cdot\|)$ is independent of N .

Remark 3. There are some intermediate cases between the two cases we presented in Theorem 3.7 and Corollary 3.8. Specifically, consider A that has an eigenvalue of 1 or -1 , which is not semisimple, but its associated Jordan block is of size \tilde{m} with $\tilde{m} < m$. That means that the power of n in the bound (13) is determined by the maximum between $m - 1$ and $2\tilde{m} - 2$. Then, the smoothness requirements from f should be modified accordingly. The details of such a proof can be easily recovered from the last section and therefore are omitted.

Theorem 3.7 shows a way to lift convergence rate results from scalars to matrices. However, two weaknesses of this approach are hidden in the constant C . First, C has as a factor C_A which consists of the condition number of Z (see (14)). This condition number can be significantly large. Second, C depends on the relation between matrix norms, which typically depends on the size of A .

Inspired by [14], we suggest an alternative approach for lifting convergence rate results from scalars to matrices, based on duality theorem for matrix norms, see e.g., [9, Chapter 3]. The bound we get is independent of the size of the matrix, but costs an additional smoothness from f . This independency means, for example, that for a given f the convergence rate for a matrix consisting of one Jordan block is the same as for a matrix consisting of many replicates of this block.

Theorem 3.9. Let $A \in \mathcal{M}_k(\mathbb{R})$ and denote by m the size of the largest Jordan block in the Jordan form of A . Assume $f \in C^{2m+2\ell}([-1, 1])$ such that $f^{(2m+2\ell)}$ is absolutely continuous and $f^{(2m+2\ell+1)}$ is of bounded variation. Then, for a given matrix norm we have

$$\|\mathcal{S}_N(f)(A) - f(A)\| \leq C \frac{1}{(N - \ell)^\ell}, \quad N > \ell,$$

where C is a constant independent of N and k . This constant can be bounded by

$$\frac{2}{\ell} \int_{-1}^1 \left\| \sum_{n=0}^{\infty} \alpha_n[f] T_n(A) T_n^{(\ell+1)}(t) \right\| dt.$$

The proof of Theorem 3.9 is given in Appendix A.4, where we lift Theorem 2.2 for matrices using an auxiliary Chebyshev operator, acting on matrix-valued functions. Similar approaches were introduced in [9, Chapter 6] and [14] for other types of matrix operators.

We conclude this section with a variant of Theorem 3.9 for δ -condense matrices.

Corollary 3.10. Let $A \in \mathcal{M}_k(\mathbb{R})$ be a δ -condense matrix and denote by m the size of the largest Jordan block in the Jordan form of A . Assume $f \in C^{m+\ell-1}([-1, 1])$ such that $f^{(m+\ell-1)}$ is absolutely continuous and $f^{(m+\ell)}$ is of bounded variation. Then, for a given matrix norm we have

$$\|\mathcal{S}_N(f)(A) - f(A)\| \leq C \frac{1}{(N - \ell)^\ell}, \quad N > \ell,$$

where C a constant independent of N and k .

4 Numerical examples

We turn to demonstrate numerically the theory developed in Section 3. In the context of numerical codes, it is worth mentioning the Chebfun software system and resources [5], where the interested reader can further study other numerical implementations and issues related to Chebyshev expansions.

This section is divided into three parts. The first part briefly explains our implementation of matrix Chebyshev expansions. The second part demonstrates the evaluation of non-analytic matrix functions by matrix Chebyshev expansions. The third part illustrates numerically the convergence properties of matrix Chebyshev expansions for Jordan block matrices.

4.1 Notes on implementation

The common practice for evaluating Chebyshev expansions starts by computing the expansion coefficients based on the discrete orthogonality property of Chebyshev polynomials. This computation is nothing but the Clenshaw-Curtis quadrature for (4), and thus can be implemented using the Fast Fourier Transform (FFT) algorithm, see e.g., [24, Chapter 5.8].

Having obtained the expansion coefficients for \mathcal{S}_N of (6), evaluating \mathcal{S}_N employs Clenshaw's algorithm (see e.g., [24, Chapter 5.5]), which exploits the three term recursion (3), as described in [24, p. 193]. Note that this algorithm is valid also for matrices since each polynomial consists only of powers of A , which means that any two such matrix polynomials of A commute. Commutativity makes scalar algorithms, such as Clenshaw's algorithm, applicable to matrices. The algorithm for evaluating matrix Chebyshev expansions is given for completeness as Algorithm 1 in Appendix C.

The entire code, of both this section and the next Section 5, is available online at <https://github.com/nirsharon/>. The numerical examples were executed on a Springdale Linux desktop equipped with 3.2 GHz i7-Intel CoreTM cpu with 16 GB of memory, and were run on Matlab 2015a.

4.2 Lifting non-analytic functions

An important motivation for using Chebyshev expansions is to calculate the matrix version of non-analytic functions. A textbook solution involving diagonalizing A might be computationally infeasible for large matrices, and numerically unstable for general matrices. The same holds if A cannot be diagonalizable, and its Jordan decomposition is used instead. Thus, we consider the matrix Chebyshev expansion as an alternative.

In our first set of examples, we use three different functions which we apply on symmetric random matrices of size 10×10 . These matrices have, with high probability, only simple eigenvalues (geometric multiplicity of one) and are diagonalizable from symmetry. We use this simple setting to demonstrate the convergence rates of Subsection 3.3.

The first function we use is

$$f_1(x) = \text{sign}(x)x^2 = \begin{cases} -x^2 & x < 0, \\ x^2 & \text{otherwise.} \end{cases} \quad (21)$$

This function is C^1 , with jump discontinuity in the second derivative at $x = 0$. Clearly, no Taylor series is available for evaluating $f_1(A)$. Since the second derivative is of bounded variation, we expect from Theorem 3.7 that the error of the truncated matrix Chebyshev expansion of order N would decay as N^{-2} (take $m = 1$ and $j = 1$). To numerically validate this theoretical result, we calculate matrix Chebyshev expansions with up to 2000 coefficients. f_1 is an odd function and thus only half of its expansion coefficients are nonzero. The error is measured using the spectral norm and given by

$$\|\mathcal{S}_N(f)(A) - f(A)\| / \left(\|A\| \|A^{-1}\| \right).$$

We present in Figure 1a the coefficients' decay, the approximation error of the truncated matrix Chebyshev expansions, and the theoretical bound (of N^{-2}) on a logarithmic scale, as functions of the expansion length N . Indeed, we see that the approximation errors fit the theoretical bound.

We repeat the last experiment with the continuous function

$$f_2(x) = \sqrt{|x|}. \quad (22)$$

The derivative of this function has a singularity at $x = 0$, nevertheless it is of bounded variation. As such, the coefficients' decay is slower than of f_1 and so we expect the convergence rate to behave like N^{-1} . As in the previous example, we show on a logarithmic scale the coefficients' decay, the convergence rate and its theoretical bound. As seen in Figure 1b, the numerical convergence rate fits its expected theoretical rate.

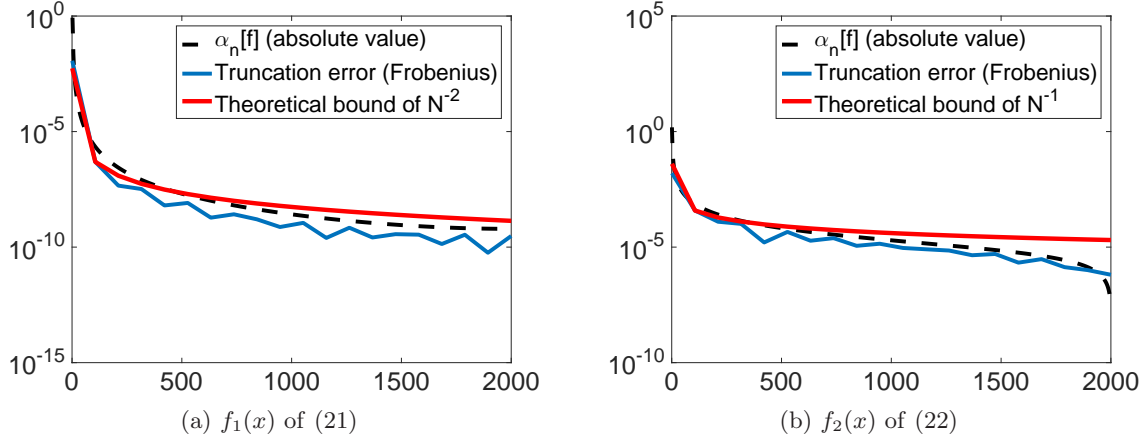


Figure 1: Lifting scalar functions to 10×10 random matrices: comparisons of the coefficients' decay, the approximation error of the truncated matrix Chebyshev expansion, and the theoretical bound of Theorem 3.7.

As a last example for this part, consider

$$f_3(x) = \frac{1}{x^2 + 0.25}. \quad (23)$$

This function is analytic around zero, but with radius of convergence of 0.5, due to its poles at $\pm 0.5i$. Therefore, the Taylor expansion of $f_3(x)$ cannot be used for matrices having eigenvalues above 0.5. The matrix that we use is of size 10×10 , with seven eigenvalues whose magnitudes larger than 0.5, as seen in

the bar plot of Figure 2a. Since f_3 has derivatives of any order everywhere on the real line, the coefficients decay rapidly and we have to use only 73 coefficients (among them only 37 nonzero) to get double precision accuracy. Figure 2b shows that indeed both the expansion coefficients and the approximation error decay rapidly.

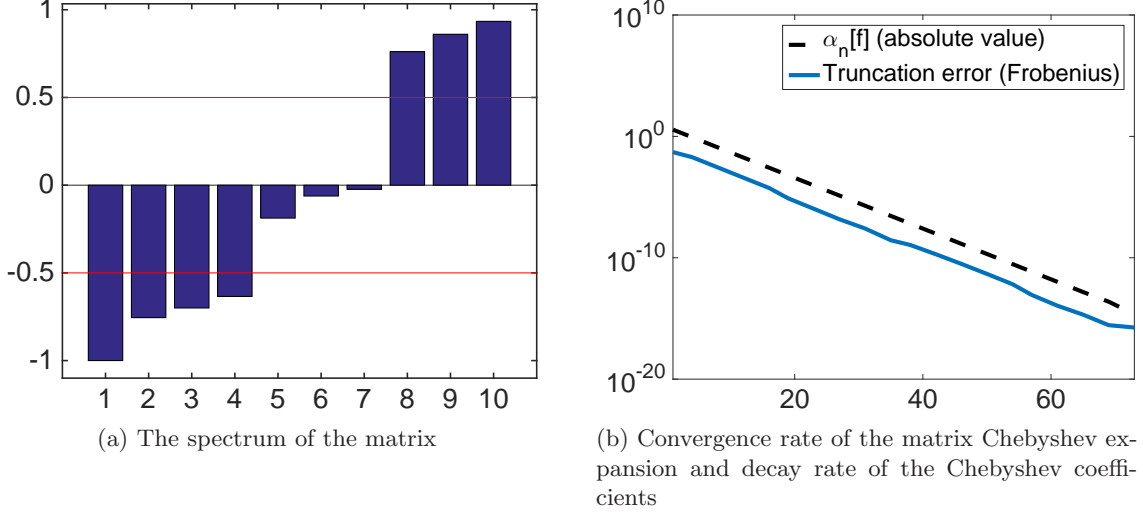


Figure 2: Evaluating f_3 of (23) on a matrix of size 10×10 .

4.3 Matrix Chebyshev expansion on Jordan blocks

As an example for applying matrix functions on non-diagonalizable matrices, we investigate the case of Jordan blocks. We focus on the two parameters of Jordan blocks: their eigenvalues and their size.

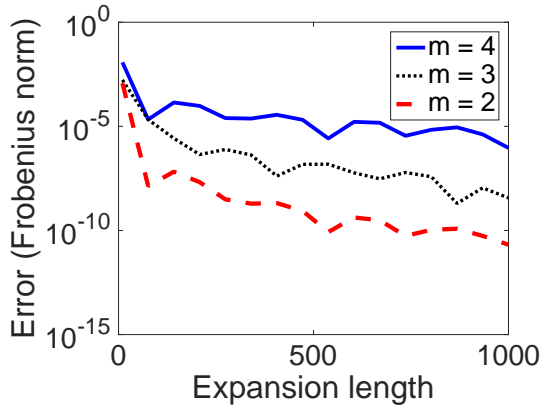
We begin by considering the function

$$f_4(x) = |x|^{3.5}, \quad x \in [-1, 1]. \quad (24)$$

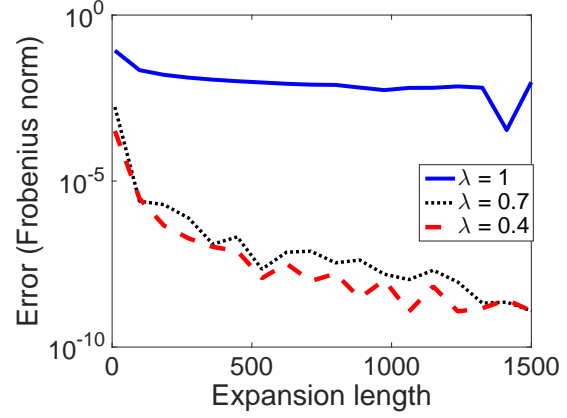
This function has 3 absolutely continuous derivatives, but the fourth one is not of bounded variation. Thus, for Jordan blocks with eigenvalue in $(-1, 1)$, Corollary 3.6 guarantees the convergence only for $m \leq 3$. We evaluate f_4 for Jordan blocks associated with the eigenvalue 0.7 and of sizes $m = 2, 3$ and 4. The expansion of f_4 for the first two Jordan blocks is guaranteed to converge, while for the third it does not. The convergence rates of the three expansions are depicted in Figure 3a, where the truncation error is given as a function of the expansion length. The error axis is in logarithmic scale, and still there is a clear difference between the decay rates of the three cases.

When considering δ -condense matrices, for f_4 and for Jordan blocks of size 3 (with eigenvalue in $[-1 + \delta, 1 - \delta]$), the convergence is guaranteed, while if the eigenvalues are 1 or -1 , the convergence is guaranteed only for Jordan blocks of size 2. We plot in Figure 3b the truncation errors for three cases of different eigenvalues 0.4, 0.7, and 1. We see that the convergence rate associated with the eigenvalue 0.7 is slightly higher than the one of 0.4, which can be explained by the growing constant (17) and Lemma 3.5. Also, the errors that are associated with the eigenvalue 1 are large and its matrix Chebyshev expansion doesn't seem to converge.

The other parameter that affects the convergence rate is the order of the largest Jordan block. In



(a) Three different sizes of Jordan blocks with eigenvalue 0.7



(b) Three Jordan blocks of order $m = 3$ correspond to the eigenvalues 0.4, 0.7, and 1

Figure 3: Truncation errors for evaluating f_4 of (24) on Jordan blocks as a function of the expansion length.

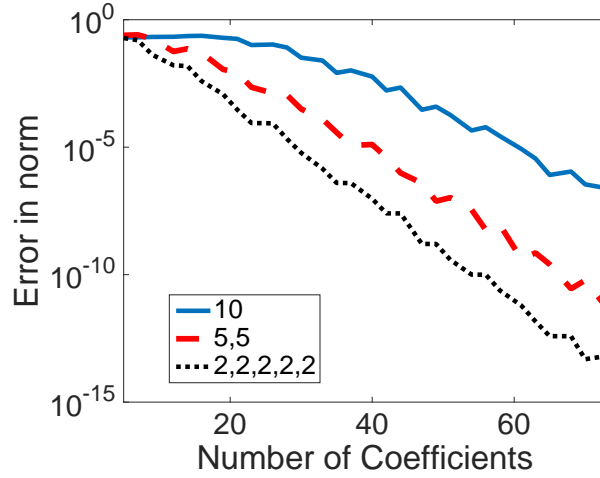


Figure 4: Convergence rates of f_3 from (23) for three different block diagonal matrices. Each block corresponds to the eigenvalue 0.5. In the legends are the different orders of the blocks on the diagonal of our test matrix.

particular, we show in Theorem 3.9 that while the size of the largest Jordan block is significant, the matrix size should not effect the convergence rate. In order to test this observation numerically, we compare three block matrices of order 10, having blocks of varying sizes on the diagonal. These blocks have the eigenvalue 0.5 on the diagonal but also 0.5 on subdiagonal entries (instead of traditionally 1). This is done to avoid high condition numbers which can obscure the results of this test. The first matrix consists of a single block of order 10×10 , the second consists of two blocks of order 5, and the third consists of five blocks, each of order 2. We use the function f_3 of (23) to allow for large Jordan blocks. The results are plotted in Figure 4, where we notice that the convergence rates depend on the size of the largest Jordan block. It is worth noting that we repeat this test with a matrix having on the diagonal two replicates of the above matrices to experience exactly identical results, which confirms numerically the observation of Theorem 3.9.

5 Application of matrix Chebyshev expansions to eigenspaces recovery

We demonstrate the possibility of using matrix Chebyshev expansions for estimating the eigenspace which corresponds to a given eigenvalue of a matrix. This problem is usually addressed by the inverse power method [7, Chapter 8], where one uses the power method on $(A - \lambda I)^{-1}$. This requires to solve at each iteration a system of linear equations of the order of the matrix. Thus, for huge matrices the conventional inverse power method might be too computationally expensive.

Recall that since matrix functions preserve similarity, the operation $f(A)$ can be interpreted as filtering the spectrum of A . Thus, we propose to use a filter function, which is designed in advance to retain eigenspaces that correspond to eigenvalues in a specific segment $I \subset \rho(A)$. The ideal filter for this task is the characteristic function of I , defined as one on I and zero otherwise. Applying such a function on A results in a lower rank matrix \tilde{A} that has, outside its kernel, only the eigenspaces of A that correspond to eigenvalues in I . Specifically, the spectrum of \tilde{A} will include only ones, for the eigenvalues in I , and zeros otherwise. Note that in order to reveal the eigenvalues themselves, we can either use as a filter the characteristic function multiplied by the identity function, or in the case of one eigenvalue, evaluate it by computing the product of any eigenvector from the recovered eigenspace with the matrix.

The main problem with evaluating the ideal filter is its poor smoothness, as it is not even continuous. Therefore, we use an alternative continuous approximation defined by

$$f(x) = \frac{1}{2} \left(1 - \operatorname{erf} \left(\frac{2}{r} (|x - c| - R) \right) \right). \quad (25)$$

Here, erf is the error function (integration over the Gaussian function), c is the center of I , R is (roughly) half of the length of I , and r controls the steepness of the transition from zero to one (the smaller r the steepest $f(x)$ is). The function $f(x)$ is smooth except at the center point c , where its first derivative has a jump discontinuity due to the absolute value. This discontinuity implies Gibbs phenomenon when approximating f with the truncated Chebyshev series. Namely, there is always a fixed magnitude of approximation error centered around c . However, approximation errors around c are almost meaningless as we are interested only in suppressing eigenspaces that correspond to eigenvalues outside I . An illustration of (25) for different sets of parameters is given in Figure 5.

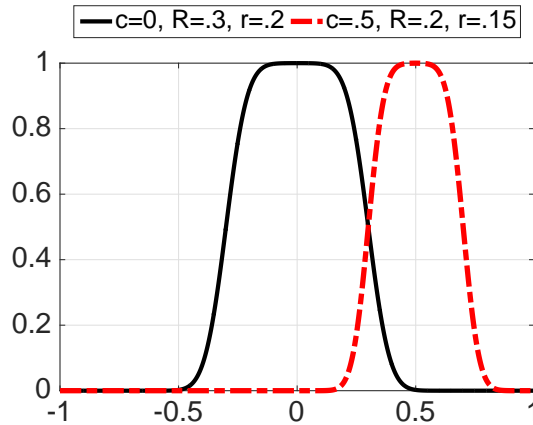


Figure 5: The filter function (25) for two different sets of parameters

We consider a scenario where A is a very large matrix, and assume we only have access to its application to any given vector. In addition we assume as apriori knowledge that we have an approximated value of the specific eigenvalue whose eigenspace we wish to recover. To make this scenario realistic, we must assume there is some gap between this eigenvalue and the rest of the spectrum, for otherwise any known variant of the power method would converge very slowly. The parameters of our filter are designed so c is the approximated eigenvalue and r and R depend on the gap. As we decrease r (corresponding to a smaller gap), the derivatives of f increase, which means we have to use more coefficients in the truncated Chebyshev series to obtain the same prescribed accuracy of approximation.

In the following we describe two examples to demonstrate the use of matrix Chebyshev expansions for the application of eigenspace recovery. In our first example, we generate symmetric matrices of varying order, up to size $15,000 \times 15,000$, which is (approximately) the maximum size of a dense matrix that fits in memory on our machine. Each matrix has a spectrum consists of ones, zeros, and halves. The goal is to reveal the 20 eigenvectors corresponding to the eigenvalue 0.5. We compare two methods for this problem: the first method is the inverse power method, implemented by solving a linear system of the form $(A - \lambda I)x = u$ with $\lambda = 0.5$ at each iteration using Matlab's implementation (the *EIG* procedure with the backslash $A \setminus b$ solver). The second method is our filtering via matrix Chebyshev expansion. To make a fair comparison between the two methods, we require the same precision of 10 digits when measuring the error by

$$\sum_{j=1}^k \|Au_j - \lambda u_j\|.$$

To reach such a precision using the matrix Chebyshev expansion, we can truncate the expansion at large N . However, that means computing many matrix-vector multiplications. A more efficient approach is to approximate $f(x)$ using some small N such as $N = 10$, and to repeat the application of this approximated function a few times. In particular, the required precision in our case is reached after seven iterations. We initial the iteration with a set of random vectors whose number is at least as large as the dimension of the subspace to be recovered. The results of the comparison between the two algorithms is presented in Figure 6, where the gap between the two method becomes significant as the size of the matrix increases.

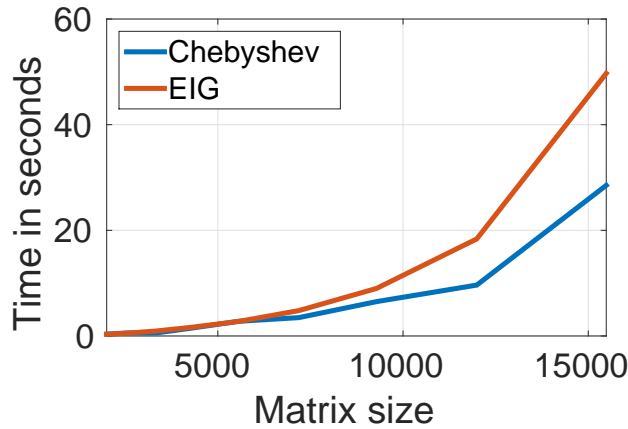


Figure 6: A timing comparison between iterative power method (by Matlab) and eigenspace recovery using matrix Chebyshev expansion. Time is measured in seconds, as a function of the matrix size.

size	50,000	100,000	500,000	1,000,000
Timing (sec.)	3.617	8.658	51.970	108.258
Precision	2.758e-10	1.660e-11	3.677e-11	5.190e-11

Table 2: Timing (in seconds) of eigenspace recovery for big matrices.

In our second experiment, we use a special type of matrices which are of the form $Q^T D Q$, where D is a diagonal matrix (the real spectrum) and Q is the orthogonal matrix of the discrete cosine transform (DCT). Namely, for a vector x , Qx is the discrete cosine transform of type II of x . The advantage of using such a matrix for our example is that realizing the vector products Qx and $Q^{-1}y = Q^T y$ for any vectors x and y is extremely fast. We examine four sizes of matrices ranging from 50,000 to 1,000,000, each matrix with a spectrum constructed as in the first example. Note that such dense, large matrices cannot be fitted into the RAM of our machine. We measure the time required for our algorithm to estimate the required eigenspace to the precision of ten digits. The results are summarized in Table 2, indicating how fast and efficient it is to use the matrix Chebyshev expansion for this task. Note that solving linear systems of these sizes using conventional methods is far from being trivial and thus applying the iterative power method is intricate.

Appendix A Complementary proofs

A.1 Proof of Lemma 2.3

Proof. By Theorem 2.2 we have that $|\alpha_n[f]| \leq \frac{C_f}{n^{2m}}$, with a positive constant C_f . By the bound (9), we have that $|\alpha_n[f]T_n^{(j)}(x)| \leq M_n = C_f \frac{n^{2j}}{(2j-1)!n^{2m}}$, for any $j = 0, 1, \dots, m-1$, and thus

$$\sum_{n=1}^{\infty} M_n = \frac{C_f}{(2j-1)!} \sum_{n=1}^{\infty} \frac{1}{n^{2(m-j)+2}} < \infty, \quad j \leq m-1. \quad (26)$$

Equation (26) guarantees that $\sum_{n=0}^{\infty} |\alpha_n[f]T_n^{(j)}(x)| < \infty$, as required.

For the second part of Lemma 2.3, we can use the Weierstrass M-test to conclude uniform convergence of

$$\sum_{n=0}^{\infty} \alpha_n[f]T_n^{(j)}(x), \quad j = 0, 1, \dots, m-1.$$

Moreover, it is clear that

$$\sum_{n=0}^N \alpha_n[f]T_n^{(j)}(x), \quad N \in \mathbb{N}, \quad j = 0, 1, \dots, m-1$$

is a polynomial of degree N and so obviously differentiable, and thus, we can differentiate it term-by-term. \square

A.2 Proof of Lemma 3.2

Proof. Decompose each Jordan block J_i as $J_i = \lambda_i I + N$, where N is the $k_i \times k_i$ nilpotent matrix

$$N = \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}.$$

The scalar matrix $\lambda_i I$ commutes with any matrix in $\mathcal{M}_{k_i}(\mathbb{R})$, and thus one can apply the Binomial formula to expand $(J_i)^n = (\lambda_i I + N)^n$, $n \in \mathbb{N}$. Consequently,

$$(J_i)^n = \begin{pmatrix} \lambda_i^n & \binom{n}{1} \lambda_i^{n-1} & \cdots & \binom{n}{k_i-1} \lambda_i^{n-k_i+1} \\ & \lambda_i^n & \ddots & \vdots \\ & & \ddots & \binom{n}{1} \lambda_i^{n-1} \\ & & & \lambda_i^n \end{pmatrix}.$$

Therefore, for $1 \leq q \leq j \leq k_i$

$$(p(J))_{q,j} = \sum_{n=0}^{\ell} a_n ((J_i)^n)_{q,j} = \sum_{n=0}^{\ell} a_n \binom{n}{j-q} \lambda_i^{n-j+q} = \frac{1}{(j-q)!} \frac{d^{j-q}}{dx^{j-q}} p(\lambda_i).$$

□

A.3 The second leading order term of $T_n^{(k)}(x)$

As noted in Remark 1, we can derive a bound on the constant in front of the n^{k-1} term of the k -th derivative $T_n^{(k)}(x)$. We denote by B_k the bound and prove by induction that $B_k \leq \frac{k(k-1)}{2}$. For the base of the induction, we get from (18) that the constants are equal to 0, 1 and 3, for $k = 1, 2$ and 3, respectively. Assume that the claim is true for any j that satisfies $0 \leq j \leq k < n$, for a fixed k . Then, for $k+1$, by (18) we get the n^k terms out of two factors: from the n^k (leading order) term of $T_n^{(k)}$ that is multiplied by $(2k-1)\nu^2$, and from the n^{k-2} term (second leading order) of $T_n^{(k-1)}$ that is multiplied by ν^2 . Combining the two factors leads to

$$B_{k+1} \leq (2k-1)\nu^{k+2} + \nu^2 B_{k-1}.$$

The latter recursion can be solved by repeatedly substituting the above bound to yield

$$B_{k+1} \leq \nu^{k+2} \sum_{j=1}^{\lfloor (k+1)/2 \rfloor} (2(k-2j)+1) = \frac{k(k+1)}{2} \nu^{k+2},$$

as required for $k+1$.

A.4 Proof of Theorem 3.9

To support the proof of Theorem 3.9 we need to define and derive some properties of an auxiliary Chebyshev term-by-term operator, acting from \mathbb{R} to $\mathcal{M}_k(\mathbb{R})$.

Definition 3. *The term-by-term Chebyshev polynomial of order N for a function $f: [-1, 1] \rightarrow \mathcal{M}_k(\mathbb{R})$ is*

$$Q_N(f)(x) = \sum_{n=0}^N \beta_n[f] T_n(x),$$

where $\beta_n[f] \in \mathcal{M}_k(\mathbb{R})$ with $(\beta_n[f])_{i,j} = \alpha_n[f(x)_{i,j}]$ for all $1 \leq i, j \leq k$.

The coefficients $\{\beta_n\}_{n=0}^N$ of Q_N are matrices defined by α_n which is the Chebyshev inner product (4), in each entry. One important relation between the Chebyshev term-by-term operator of Definition 3 and the truncated Chebyshev series for scalar functions (6) is as follows.

Lemma A.1. *Let $f: [-1, 1] \rightarrow \mathcal{M}_k(\mathbb{R})$ be a matrix-valued function such that $f(x)_{i,j} \in C([-1, 1])$, for all $1 \leq i, j \leq k$. Then, for any fixed $B \in \mathcal{M}_k(\mathbb{R})$*

$$\alpha_n \left[\text{tr} \left(f(t) B^T \right) \right] = \text{tr} \left(\beta_n[f(t)] B^T \right).$$

In words, the trace and the (Chebyshev) inner product commute.

Proof. Define $h: [-1, 1] \rightarrow \mathbb{R}$ by $h(t) = \text{tr} \left(f(t) B^T \right)$. The trace operator is linear and since f is continuous in each entry so is h . Therefore, $\alpha_n[h]$ is well-defined and

$$\begin{aligned} \alpha_n[h] &= \alpha_n \left[\text{tr} \left(f(t) B^T \right) \right] = \frac{2}{\pi} \int_{-1}^1 \frac{\sum_{i,j=1}^k f(t)_{i,j} B_{i,j} T_n(t)}{\sqrt{1-t^2}} dt \\ &= \sum_{i,j=1}^k B_{i,j} \frac{2}{\pi} \int_{-1}^1 \frac{f(t)_{i,j} T_n(t)}{\sqrt{1-t^2}} dt \\ &= \sum_{i,j=1}^k B_{i,j} \alpha_n[f(t)_{i,j}] = \text{tr} \left(\beta_n[f(t)] B^T \right). \end{aligned}$$

□

Equipped with Lemma A.1, we are ready for the first step, where we lift Theorem 2.2 to the case of the Chebyshev term-by-term operator for matrix-valued functions. To this end, we use fundamentals from matrix duality theorem, stating that for any $X \in \mathcal{M}_k(\mathbb{R})$ and a matrix norm $\|\cdot\|$, there exists a matrix Y such that $\|X\| = \text{tr}(XY^T)$ and $\|Y\|^D = 1$, where $\|\cdot\|^D$ is the dual norm defined as

$$\|X\|^D = \max_{\|Y\| \leq 1} \{\text{tr}(XY^T)\}.$$

The duality ensures that $(\|\cdot\|^D)^D = \|\cdot\|$ and that $|\text{tr}(XY^T)| \leq \|X\| \|Y\|^D$. For more details see [10, Chapter 5].

Theorem A.2. Let $f: [-1, 1] \rightarrow \mathcal{M}_k(\mathbb{R})$ be such that $f_{i,j} \in C^{m-1}([-1, 1])$ where $f_{i,j}^{(m-1)}$ is absolutely continuous and $f_{i,j}^{(m)}$ is of bounded variation for any $1 \leq i, j \leq k$. Then,

$$\|f(x) - Q_N(f)(x)\| \leq C \frac{1}{(N-m)^m}, \quad N > m,$$

where C is a constant independent of x and N .

Proof. Denote by $E_N(Q, f)(x) = f(x) - Q_N(f)(x)$ the error matrix at x . By the duality theorem for matrix norms, for any given $A \in \mathcal{M}_k(\mathbb{R})$ and a matrix norm $\|\cdot\|$, there exists a matrix B with $\|B\|^D = 1$ such that $\|A\| = \text{tr}(AB^T)$. Now, let $x_0 \in [-1, 1]$ be a fixed point and let B_0 be the corresponding matrix such that

$$\|E_N(Q, f)(x_0)\| = \text{tr} \left(E_N(Q, f)(x_0) B_0^T \right) = \text{tr} \left(f(x_0) B_0^T \right) - \text{tr} \left(Q_N(f)(x_0) B_0^T \right).$$

In other words, $\|E_N(Q, f)(x_0)\| = h(x_0) - \text{tr} \left(\sum_{n=0}^{N'} \beta_n [f] T_n(x_0) B_0^T \right)$ where $h(t) = \text{tr} \left(f(t) B_0^T \right)$ (as in the proof of Lemma A.1). Note that $T_n(x_0)$, $0 \leq n \leq N$, are scalars, and by the linearity of the trace operator

$$\text{tr} \left(\sum_{n=0}^{N'} \beta_n [f] T_n(x_0) B_0^T \right) = \sum_{n=0}^{N'} T_n(x_0) \text{tr} \left(\beta_n [f] B_0^T \right). \quad (27)$$

By Lemma A.1, the sum on the right-hand-side of (27) is equal to $\sum_{n=0}^{N'} \alpha_n \left[\text{tr} \left(f(t) B_0^T \right) \right] T_n(x_0)$, which means that

$$\|E_N(Q, f)(x_0)\| = h(x_0) - \sum_{n=0}^{N'} \alpha_n [h] T_n(x_0). \quad (28)$$

The right hand side of (28) is the Chebyshev error term of the scalar function h , and therefore, to bound it using Theorem 2.2, we need to verify that $h(x)$ satisfies the conditions of that theorem. The regularity of h is directly inherited from the components of f , and so

$$h^{(j)}(x) = \text{tr} \left(f^{(j)}(x) B_0^T \right), \quad j = 1, \dots, m.$$

Thus, it remains to verify the finiteness of $\|h^{(m)}\|_{TV}$. By the duality theorem, $|\text{tr}(AB^T)| \leq \|A\| \|B\|^D$, where in our case $\|B_0^T\|^D = 1$, and consequently,

$$\|h^{(m)}\|_{TV} \leq \int_{-1}^1 \|f^{(m+1)}(t)\| \|B_0^T\|^D dt \leq \int_{-1}^1 \|f^{(m+1)}(t)\| dt < \infty.$$

The last inequality can be verify easily using the max norm and the conditions on $f_{i,j}$. Applying Theorem 2.2 to h leads to the required bound due to (28). \square

There are two important notes regarding Theorem A.2. First, the constant C of the bound is independent of k (the order of the matrix). Nevertheless, the size of the largest Jordan form does have an effect and C can be bounded by $\frac{2}{m} \int_{-1}^1 \|f^{(m+1)}(t)\| dt$. Second, there are no restrictions on the matrix

norm, except of being sub-multiplicative. We preserve these properties also in Theorem 3.9.

Proof of Theorem 3.9. Let $g: [-1, 1] \rightarrow \mathcal{M}_k(\mathbb{R})$ be a matrix-valued function defined by

$$g(x) = \sum_{n=0}^{\infty} \alpha_n[f] T_n(A) T_n(x). \quad (29)$$

Note that Theorem 3.3 guarantees absolute convergence of the matrix Chebyshev expansion of f , and thus, g is well-defined since

$$\sum_{n=0}^{\infty} \|\alpha_n[f] T_n(A) T_n(x)\| = \sum_{n=0}^{\infty} \|\alpha_n[f] T_n(A)\| |T_n(x)| \leq \sum_{n=0}^{\infty} \|\alpha_n[f] T_n(A)\| < \infty.$$

On the one hand, $g(1) = f(A)$ since $T_n(1) = 1$ for all $n = 0, 1, 2, \dots$. On the other hand, using the operator of Definition 3

$$Q_N(g)(1) = \sum_{n=0}^N \beta_n[g] T_n(1) = \sum_{n=0}^N \beta_n[g], \quad (30)$$

where each matrix $\beta_n[g]$ is defined as

$$\beta_n[g]_{i,j} = \frac{2}{\pi} \int_{-1}^1 \frac{(g(t))_{i,j} T_n(t)}{\sqrt{1-t^2}} dt = \frac{2}{\pi} \int_{-1}^1 \frac{(\sum_{\ell=0}^{\infty} \alpha_{\ell}[f] T_{\ell}(A) T_{\ell}(t))_{i,j} T_n(t)}{\sqrt{1-t^2}} dt.$$

Each element of $\{\sum_{\ell=0}^p \alpha_{\ell}[f] T_{\ell}(A)\}_{p=1}^{\infty}$ is bounded by $\sum_{\ell=0}^{\infty} \|\alpha_{\ell}[f] T_{\ell}(A)\|$, and thus, we can use the bounded convergence theorem to interchange the integral and sum. Therefore, we have

$$\beta_n[g]_{i,j} = \frac{2}{\pi} \sum_{\ell=0}^{\infty} \int_{-1}^1 \frac{(\alpha_{\ell}[f] T_{\ell}(A) T_{\ell}(t))_{i,j} T_n(t)}{\sqrt{1-t^2}} dt.$$

Note that $T_{\ell}(t)$ is a scalar and $\alpha_{\ell}[f] T_{\ell}(A)$ is independent of the integration variable. Thus, we can rearrange the latter equation and use the orthogonality of the Chebyshev polynomials to get

$$\begin{aligned} \beta_n[g]_{i,j} &= \frac{2}{\pi} \sum_{\ell=0}^{\infty} (\alpha_{\ell}[f] T_{\ell}(A))_{i,j} \int_{-1}^1 \frac{T_{\ell}(t) T_n(t)}{\sqrt{1-t^2}} dt \\ &= \sum_{\ell=0}^{\infty} (\alpha_{\ell}[f] T_{\ell}(A))_{i,j} \delta_{\ell,n} = (\alpha_n[f] T_n(A))_{i,j}, \end{aligned}$$

where $\delta_{\ell,n}$ is the Kronecker delta. Combining the latter equation with (30) we get that $Q_N(g)(1) = \sum_{n=0}^N \alpha_n[f] T_n(A) = S_N(f)(A)$ and therefore,

$$\|g(1) - Q_N(g)(1)\| = \|f(A) - S_N(f)(A)\|. \quad (31)$$

It remains to show that the assumptions on f implies that $g \in C^{\ell+1}$ so that we can apply Theorem A.2 on g to get the error bound $\frac{C}{(N-\ell)^{\ell}}$. This is done by considering the bound $|T_n^{(j)}(x)| \leq n^{2j}$, and the bound

(13) on $\|T_n(A)\|$. Then, we have

$$\sum_{n=1}^{\infty} \left\| \alpha_n[f] T_n(A) T_n^{(j)}(x) \right\| < \infty, \quad 0 \leq j \leq \ell + 1,$$

and Lemma 2.3 implies the required smoothness. \square

Appendix B Matrix functions

Given $f: \mathbb{R} \rightarrow \mathbb{R}$ we want to evaluate the matrix function $f(A)$, that is, to lift the scalar function f to $f: \mathcal{M}_k(\mathbb{R}) \rightarrow \mathcal{M}_k(\mathbb{R})$ where $\mathcal{M}_k(\mathbb{R})$ is the set of $k \times k$ real matrices having real spectrum. We next provide two standard definitions for matrix functions. The interested reader is referred to [8, Chapter 1] for more details as well as for additional equivalent definitions.

Definition 4 (Matrix function via Taylor series). *Let $A \in \mathcal{M}_k(\mathbb{R})$, and assume f is an analytic function with radius of convergence exceeding $\rho(A)$. Then, by Taylor's formula, one defines*

$$f(A) = \sum_{n=0}^{\infty} \frac{f^{(n)}(0)}{n!} A^n. \quad (32)$$

Note that the absolute convergence of (32) is guaranteed under the assumptions of Definition 4.

Definition 5 (Matrix function via Jordan form). *Assume $A \in \mathcal{M}_k(\mathbb{R})$ has the Jordan form $A = Z^{-1}JZ$, where J is a block diagonal matrix. Denote by J_i the Jordan block of order $k_i \times k_i$, corresponding to the eigenvalue λ_i , namely*

$$J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix}. \quad (33)$$

Also, denote by $m = \max_i k_i$ the largest block size of J . Then, for f having at least $m - 1$ derivatives on $[-1, 1]$ one defines

$$f(A) = Z^{-1} f(J) Z,$$

where $f(J)$ is a block diagonal matrix, whose i -th block $f(J_i)$ is an upper triangular matrix, given by

$$f(J_i) = \begin{pmatrix} f(\lambda_i) & f'(\lambda_i) & \cdots & \frac{f^{(k_i-1)}(\lambda_i)}{(k_i-1)!} \\ & f(\lambda_i) & \ddots & \vdots \\ & & \ddots & f'(\lambda_i) \\ & & & f(\lambda_i) \end{pmatrix}. \quad (34)$$

Definitions 4 and 5 coincide when the conditions of Definition 4 hold. In addition, there are several other equivalent standard definitions for $f(A)$, based for example, on Hermite interpolation of the roots of the minimal polynomial of A , see e.g. [8, Chapter 1].

Both Definitions 4 and 5 have inherent computational drawbacks. In particular,

1. Definition 4 requires evaluating high order derivatives of the given function. In addition, requiring f to be analytic is in some cases too restrictive.
2. Definition 5 is suitable for non-analytic functions, but requires the Jordan form of A whose computation may be unstable due to the condition number of Z , see e.g., [7, Chapter 7].

It is worth noting that most of the other equivalent standard definitions of a matrix function are also difficult to implement numerically (for an arbitrary given function). For example, evaluating the minimal polynomial of a matrix is usually avoided. This reasoning encourages us to give an alternative, equivalent approach.

Appendix C Clenshaw's algorithm for matrices

Algorithm 1 Clenshaw's algorithm for evaluating a truncated matrix Chebyshev expansion

Input: Matrix A of order k and coefficients $\{c_i\}_{i=0}^N$.

Output: Truncated Chebyshev expansion $\sum_{n=0}^N c_n T_n(A)$.

```

1:  $T \leftarrow 2 \cdot A - I_k$ 
2:  $d \leftarrow 0$ 
3:  $dd \leftarrow 0$ 
4: for  $n \leftarrow N$  downto 2 do
5:    $b \leftarrow d$ 
6:    $d \leftarrow 2 \cdot T \cdot d - dd + c_n \cdot I_k$ 
7:    $dd \leftarrow b$ 
8: end for
9: return  $T \cdot d - dd + 0.5 \cdot c_0 \cdot I_k$ 
```

References

- [1] Luca Bergamaschi and Marco Vianello. Efficient computation of the exponential operator for large, sparse, symmetric matrices. *Numerical Linear Algebra with Applications*, 7(1):27–45, 2000.
- [2] Serge Bernstein. Quelques remarques sur l'interpolation. *Mathematische Annalen*, 79(1):1–12, 1918.
- [3] Claude R Dietrich and Garry N Newsam. Efficient generation of conditional simulations by Chebyshev matrix polynomial approximations to the symmetric square root of the covariance matrix. *Mathematical geology*, 27(2):207–228, 1995.
- [4] Eid H Doha. The coefficients of differentiated expansions and derivatives of ultraspherical polynomials. *Computers & Mathematics with Applications*, 21(2):115–122, 1991.
- [5] Tobin A Driscoll, Nicholas Hale, and Lloyd N Trefethen. Chebfun guide, 2014.

- [6] Vladimir L Druskin and Leonid A Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *USSR Computational Mathematics and Mathematical Physics*, 29(6):112–121, 1989.
- [7] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.
- [8] Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.
- [9] Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge University Press, 1991.
- [10] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge University Press, 2012.
- [11] Benjamin Lepson. Upper bounds for Dirichlet kernels and for Tchebycheff polynomials of the second kind. *Journal of Mathematical Analysis and Applications*, 48(1):139–149, 1974.
- [12] Jörg Liesen and Petr Tichý. On best approximations of polynomials in matrices in the matrix 2-norm. *SIAM Journal on Matrix Analysis and Applications*, 31(2):853–863, 2009.
- [13] John C Mason and David C Handscomb. *Chebyshev polynomials*. CRC Press, 2010.
- [14] Roy Mathias. Approximation of matrix-valued functions. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1061–1063, 1993.
- [15] Victor May, Yosi Keller, Nir Sharon, and Yoel Shkolnisky. An algorithm for improving non-local means operators via low-rank approximation. *IEEE Transactions on Image Processing*, 25(3):1340–1353, 2016.
- [16] Paolo Novati and Igor Moret. The computation of functions of matrices by truncated faber series. *Numerical functional analysis and optimization*, 22(5):697–720, 2001.
- [17] Kelley R Pace and James P LeSage. Chebyshev approximation of log-determinants of spatial weight matrices. *Computational Statistics & Data Analysis*, 45(2):179–196, 2004.
- [18] Michael J D Powell. On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria. *The Computer Journal*, 9(4):404–407, 1967.
- [19] Theodore J Rivlin. *The Chebyshev polynomials*. Wiley, New York, 1974.
- [20] Hillel Tal-Ezer. Polynomial approximation of functions of matrices and applications. *Journal of Scientific Computing*, 4(1):25–60, 1989.
- [21] Hillel Tal-Ezer and R Kosloff. An accurate and efficient scheme for propagating the time dependent Schrödinger equation. *The Journal of chemical physics*, 81(9):3967–3971, 1984.
- [22] Lloyd N Trefethen. *Spectral methods in MATLAB*. SIAM, 2000.
- [23] Lloyd N Trefethen. *Approximation theory and approximation practice*. SIAM, 2013.
- [24] William T Vetterling, Saul A Teukolsky, William H Press, and Brian P Flannery. *Numerical recipes example book (C)*. JSTOR, 1992.